Exploring Faster Street-Level Semantic Segmentation with Learnable Resized and Pixel-Shuffled MobileNets

Miguel Luis R. Lagahit [1, 2, 3, *], Matthias S. Wilhelm [4], Masashi Matsuoka [1, 2]

[1] Department of Architecture and Building Engineering, Institute of Science Tokyo, Japan - {lagahit.m.297e, matsuoka.m.e594}@m.isct.ac.jp
[2] Academy of Super Smart Society, Institute of Science Tokyo, Japan
[3] Intelligent Platforms Research Institute, National Institute of Advanced Industrial Science and Technology, Japan
[4] Dynamic Map Platform Co., Ltd., Japan - matthias.wilhelm@dynamic-maps.co.jp

Keywords: Street-Level Imagery, Semantic Segmentation, Lightweight Deep Learning Model.

Abstract

This initial exploratory study investigates enhancements for real-time urban segmentation by integrating prominent scaling techniques, such as learnable resizing and pixel shuffling, into MobileNets. Using the Cityscapes dataset, we evaluate MobileNetV3 and MobileNetV4 to achieve a balance between computational efficiency and segmentation accuracy, particularly for high-resolution street-level imagery. A common way to proceed is to make use of traditional resizing methods to reduce image size and improve inference speed, but they often degrade important details, leading to lower segmentation accuracy. To address this, we incorporate a learnable resizer that optimizes downsampling while preserving critical features, along with pixel shuffling to efficiently restore spatial details during upsampling. Our results indicate that integrating learnable resizing and pixel shuffling improved segmentation accuracy by 9-14% compared to traditional resizing, and increased speed by 36-50% relative to no resizing. We also observed that MobileNetV4 continued to surpass MobileNetV3 in accuracy. Overall, the learnable resizer significantly mitigates accuracy loss due to downsampling, while pixel shuffling improves segmentation consistency with minimal impact on speed. These enhancements allow for better preservation of fine details, reducing misclassifications in complex urban scenes. By optimizing MobileNets with learnable resizing and upsampling techniques, we provide a practical solution for resource-constrained environments, such as mobile and edge computing platforms. This approach of combining efficiency-boosting techniques and lightweight architectures enables fast and accurate segmentation for urban and environmental monitoring, improving deep learning model performance in real-world applications without sacrificing speed or accuracy.

1. Introduction

1.1 Introduction

MobileNets are lightweight deep learning models designed fast, accurate inference on resource-constrained devices such as mobile phones and edge computing platforms. They are advantageous for real-time applications, such as quickly identifying different parts of an image to better understand the surroundings for tasks such as urban and environmental monitoring. Recent versions, MobileNetV3 (Howard et al., 2019) and MobileNetV4 (Qin et al., 2024), introduced improvements that further enhance the model's predictive speed and accuracy.

A key challenge in segmentation tasks, especially when working with street-level imagery, is processing high spatial resolution images while maintaining real-time performance. Since models learn details to distinguish between classes in an image, high-resolution inputs are often required. However, processing such large images takes more time, significantly slowing down inference speed. A common solution is to reduce the scale of the image before using it as input to a deep learning model. While this reduces computational costs, it can lead to the loss of details, making it harder for the model to correctly identify which class features belong to, negatively impacting segmentation accuracy.

As an initial exploratory step toward improving this trade-off, we investigate a hybrid approach that integrates learnable resizing (Talebi et al., 2021) to optimize downsampling and pixel shuffling (Shi et al., 2016) for efficient upsampling. The learnable resizer aims to preserve key features while reducing image scale, and pixel shuffling reorganizes feature maps to reconstruct spatial details in the upscaled output. This exploratory framework is designed to test the feasibility and practicality of incorporating resizing and upsampling strategies based on deep learning into lightweight MobileNet models. Although not all-inclusive, the approach lays fundamental insights for further investigation and improvement, particularly in real-world, resource-limited deployment scenarios, aiming to balance computational efficiency and segmentation accuracy.

2. Methodology

2.1 Dataset

We used the Cityscapes dataset, a widely recognized large-scale benchmark for urban scene understanding (Cordts et al., 2016). The dataset consists of high-resolution street-level images captured from various cities, providing a diverse representation of the urban environment. It includes pixel-level annotations across multiple classes.

For this study, we used a subset of the Cityscapes dataset, consisting of 3,475 finely labelled images from their training set. To simplify the segmentation task, we use eight broader categorized classes, such as flat, objects, vehicles. This label grouping helps the model better distinguish key urban features.

To ensure a balanced evaluation, we split the dataset into 17% for training, 3% for validation, and 80% for testing. This setup ensures sufficient samples for training to learn meaningful patterns while reserving a small validation set to aid in preventing overfitting. Additionally, maintaining a sufficiently large testing dataset allows for a comprehensive performance assessment, ensuring that the model generalizes well to new, unseen urban scenes.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W4-2025-293-2026 | © Author(s) 2026. CC BY 4.0 License.

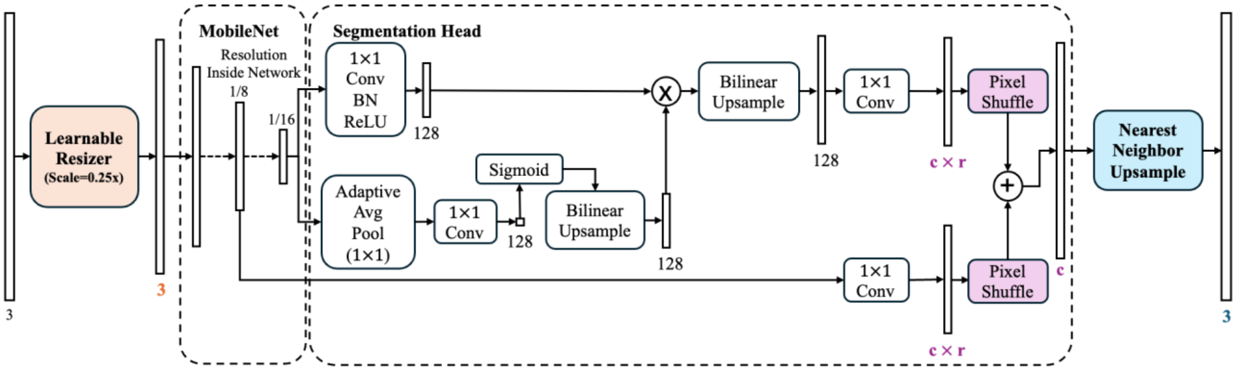

Figure 1. MobileNet enhanced with a learnable resizer in the input and pixel shuffling in the Lite R-ASPP segmentation head.

2.2 MobileNets

MobileNets are deep learning models designed for computationally constrained devices, like mobile and edge devices. Despite being lightweight, they still offer strong performance in making predictions. They are optimized to deliver fast inference speeds to meet the demands of real-time segmentation.

MobileNetV3 introduces several structural novelties that enhance both efficiency and accuracy. One key improvement is the use of inverted bottleneck layers coupled with a squeeze-and-excitation (SE) weighting approach. Another significant contribution is the incorporation of Lite Reduced Atrous Spatial Pyramid Pooling (LR-ASPP), which further accelerates segmentation. (Howard et al., 2019)

Building on these improvements, MobileNetV4 introduces more structural changes in the network, one of which is the Universal Inverted Bottleneck (UIB), which considers several combinations and permutations of widely known successful convolutional operations to further enhance accuracy and efficiency. Additionally, MobileNetV4 incorporates multi-head query attention (MQA) modules in its hybrid versions to benefit from transformer-based networks, particularly their ability to capture broader contextual information. (Qin et al., 2024) These combined improvements make MobileNetV4 particularly even more effective for tasks that need quick segmentation.

However, since MobileNetV4 is primarily designed for object detection, we integrate its backbone with the segmentation head of MobileNetV3. This combination aims to balances efficiency and accuracy, ensuring robust feature extraction for dense pixel-wise predictions. In this study, both models will be tested to evaluate their effectiveness.

2.3 Learnable Resizer

Traditional image resizing methods, such as bilinear or bicubic interpolation, are widely used to scale images while retaining a visually appealing representation. However, these methods are designed for human perception rather than preserving details crucial for deep learning models. As a result, they fail to retain features necessary for tasks like semantic segmentation.

To address this limitation, the learnable resizer couples traditional image resizers with learnable CNN operations to improve resizing performance for deep learning-based tasks. Unlike traditional resizing methods, learnable resizers produce scaled images that appear distorted, but these distortions are intentionally created to improve downstream performance. Through joint training with a deep learning model, the learnable resizer can continually adapt to maximize its performance by retaining important information lost in the resizing operation. (Talebi et al., 2021)

In our approach, we introduce a learnable resizer before the network, downsampling input images to one-quarter of their original size to preserve critical features in the segmentation task. This significantly reduces inference time while minimizing accuracy loss, making the approach better suited for real-time applications.

2.4 Pixel Shuffling

Pixel Shuffle is a computationally efficient upsampling method originally used to achieve super resolution. It offers a new structured way to increase the scale of an image while preserving details. It achieves this by reorganizing feature map channels into spatial dimensions, transforming an input of shape H × W × (C·scale^2) into (scale·H) × (scale·W) × C. This effectively distributes learned features and reduces the slowness associated with conventional transposed convolutions at minimal computational cost. (Shi et al., 2016)

In our implementation, we integrate pixel shuffling into the segmentation head to efficiently upsample the output to its original resolution. Since our proposed approach involves an initial downsampling before an image is fed to the network, we rely on pixel shuffle to restore the output of the network back to its original spatial resolution. By leveraging this technique, we attempt to preserve details while maintaining computational efficiency, ensuring high-quality segmentation results even when working with downsampled inputs.

2.5 Evaluation

Model performance is assessed using two key metrics: F1-score and inference speed. The F1-score balances precision and recall, providing a reliable measure of segmentation accuracy, while the inference speed determines the potential for real-time applicability. To account for model uncertainty, segmentation scores were computed per image, and we report the mean and standard deviation across the dataset.

To enhance segmentation consistency, we consolidated certain semantic categories that are difficult to separate at lower scales. Specifically, the construction class are grouped with the object class, while the vehicle class are merged with the human class. This helps the model focus on recognizing broad structural feature patterns. By adopting this approach, we ensure that the segmentation model prioritizes larger, more prominent regions. This aligns with our goal of optimizing performance for real-world applications.

Method	Classification F1-Score (%)							Inference Speed (sec)
	Void	Flat	Const. & Object	Nature	Sky	Human & Vehicle	Mean	
MobileNetV3	78.8±12.1	95.8±07.4	87.3±10.1	88.3±10.2	88.7±14.5	84.4±12.3	87.1±05.9	0.75
MobileNetV3 0.25x	38.8±07.3	86.9±08.8	75.3±14.1	78.5±17.1	79.6±20.4	61.6±18.9	69.7±07.8	0.10
MobileNetV3 0.25x +LR	67.9±11.8	93.4±07.0	78.8±15.5	78.5±15.1	82.2±15.0	67.9±19.1	77.9±06.9	0.37
MobileNetV3 0.25x +PS	73.1±12.7	92.5±07.2	71.8±19.8	68.9±21.0	72.6±20.4	55.7±23.0	72.4±08.2	0.12
MobileNetV3 0.25x +LR +PS	79.4±13.7	94.4±07.9	78.4±17.0	77.2±17.6	81.3±18.8	67.9±20.4	78.8±12.1	0.38
MobileNetV4	82.8±11.6	96.8±07.1	90.0±08.4	90.1±09.2	90.7±12.2	89.1±09.3	89.8±05.0	1.19
MobileNetV4 0.25x	33.4±07.2	86.7±09.1	76.5±13.2	80.5±14.5	80.1±20.5	67.8±16.5	70.4±07.4	0.17
MobileNetV4 0.25x +LR	69.9±12.1	94.2±07.1	81.4±14.5	81.9±12.9	84.1±12.9	74.7±16.6	80.9±06.3	0.43
MobileNetV4 0.25x +PS	73.2±12.7	93.4±07.5	73.9±19.2	73.4±18.9	77.6±18.5	62.7±22.0	75.6±07.9	0.17
MobileNetV4 0.25x +LR +PS	82.7±13.6	95.6±07.7	82.9±14.4	83.4±14.4	87.6±15.5	76.2±17.1	84.6±07.1	0.44

Table 1. An overview of model performance. (0.25x means that the image has been downsampled.)

3. Results and Discussion

Table 1 presents an overview of the experimental results, comparing the performance of MobileNetV3, MobileNetV4, and our proposed improvements in terms of segmentation accuracy and inference speed. MobileNetV3 and MobileNetV4 showed improvements of approximately 9% and 14% in mean segmentation accuracy, respectively. This suggests minor enhancements in feature extraction with the newer version in the subset of the Cityscapes Dataset.

Analyzing the proposed scaling techniques relative to traditional resizing show that both learnable resizing and pixel shuffling independently improved overall segmentation performance. Individually, learnable resizing contributed an additional 8-9% improvement in accuracy, while pixel added 3-5%. Although learnable resizing had better accuracy gains, pixel shuffling had a stronger impact on inference speed, reducing it by up to 84-86% compared to 51-63% for learnable resizing. When combined, inference speed slowed by 2.5x to 3.2x compared to simply using pixel shuffling alone. These results highlight a clear trade-off pixel shuffling is sufficient when speed is prioritized, whereas combining both methods is preferable when accuracy is more critical, and a slight additional latency is acceptable.

Our initial observation suggests that MobileNetV3 remains better than MobileNetV4 due to a small difference of less than 1% on the 0.25x scale. However, this trend shifted with the incorporation of the learnable resizer. With this integration, MobileNetV4 achieved an 11% increase in segmentation accuracy while maintaining inference speeds below half a second. As a result, MobileNetV4 not only outperformed MobileNetV3 by 2.2% in accuracy but also reduced the gap in inference speeds was reduced to less than 0.1 seconds.

Further improvements were achieved with the addition of pixel shuffling in the segmentation head. This restored the segmentation output back to its original size while increasing segmentation accuracy by 1-4% for both models, demonstrating its capability to effectively preserve details upon upsamling. More importantly, this improvement came at a minimal cost, adding only 0.01 seconds to the inference speed. By applying pixel shuffling, the segmentation difference between the original and downsampled was further reduced by by 5–8 while keeping inference speeds below half a second. These results collectively indicate that combining learnable resizing and pixel shuffling can substantially improve both the accuracy and consistency of MobileNets without compromising their real-time performance.

Method	Mean Recall (%)	Mean Precision (%)
MobileNetV3	85.3±08.8	85.6±08.4
MobileNetV3 ¼	72.7±08.8	69.4±09.8
MobileNetV3 ¼ +LR	75.3±09.7	77.0±08.9
MobileNetV3 ¼ +PS	69.6±10.1	72.6±09.2
MobileNetV3 ¼ +LR +PS	77.5±10.1	79.2±09.5
MobileNetV4	87.8±08.4	88.1±07.9
MobileNetV4 ¼	71.9±08.8	74.3±09.6
MobileNetV4 ¼ +LR	78.0±09.2	80.3±08.6
MobileNetV3 ¼ +PS	73.2±10.2	75.6±07.9
MobileNetV4 ¼ +LR +PS	82.5±09.6	83.9±08.9

Table 2. Segmentation performance.

Examining Table 2, which details precision and recall values of the segmentation results, reveals that MobileNets consistently maintain a stable balance between the two metrics. The minimal difference in precision and recall indicates that the models are effectively identifying positive cases while keeping false positives to a minimum, ensuring reliable performance. This stability remains even after the addition of learnable resizing and pixel shuffling, demonstrating that these complement MobileNets without disrupting their predictive balance. Furthermore, learnable resizing independently introduces less discrepancy between precision and recall compared to pixel shuffling, highlighting its stronger contribution to maintaining predictive consistency. These results confirm that MobileNets, when combined with resizing and upsampling techniques, can deliver accurate and fast segmentation results for real-time applications.

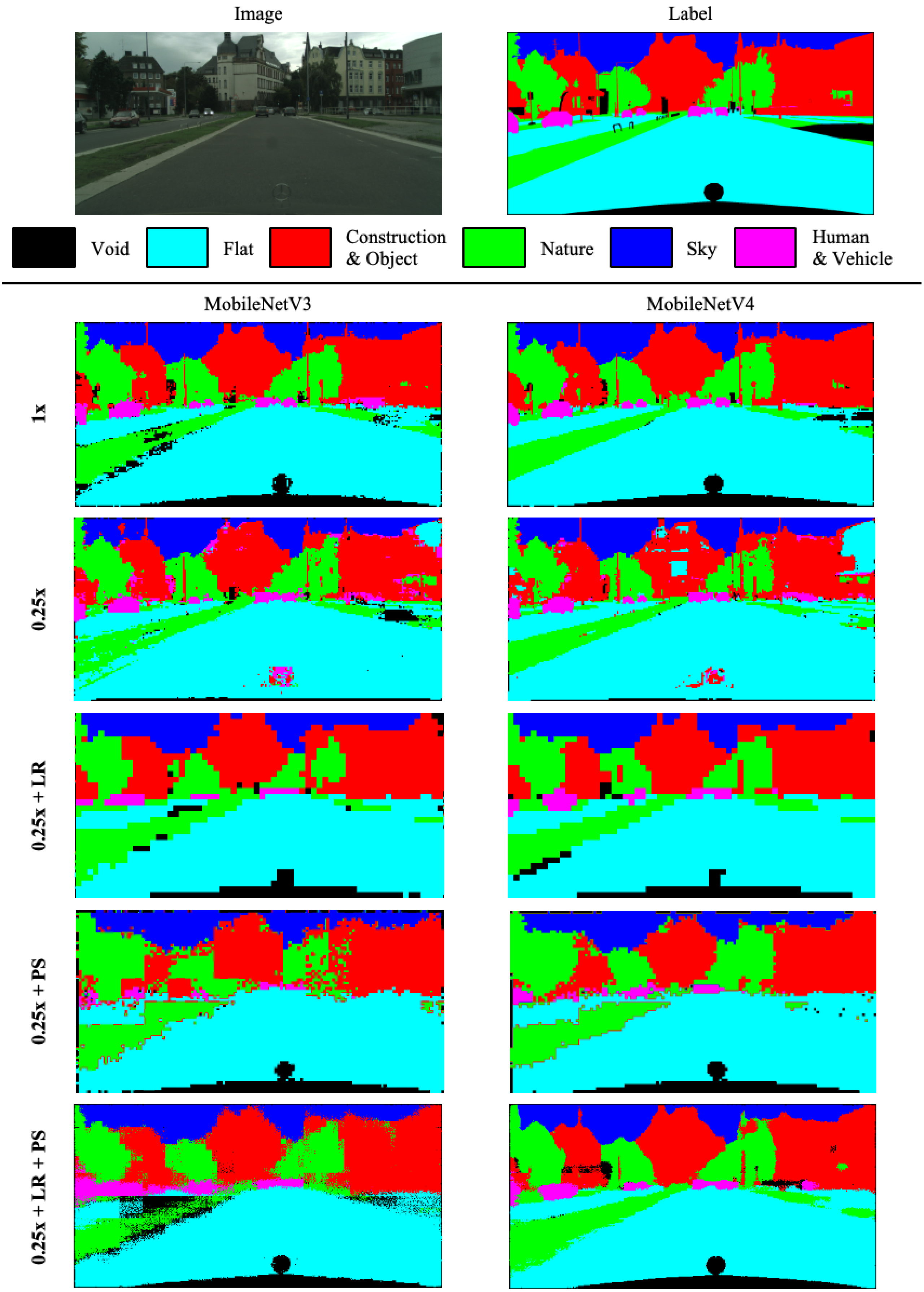

Figure 2. Sample segmentation results.

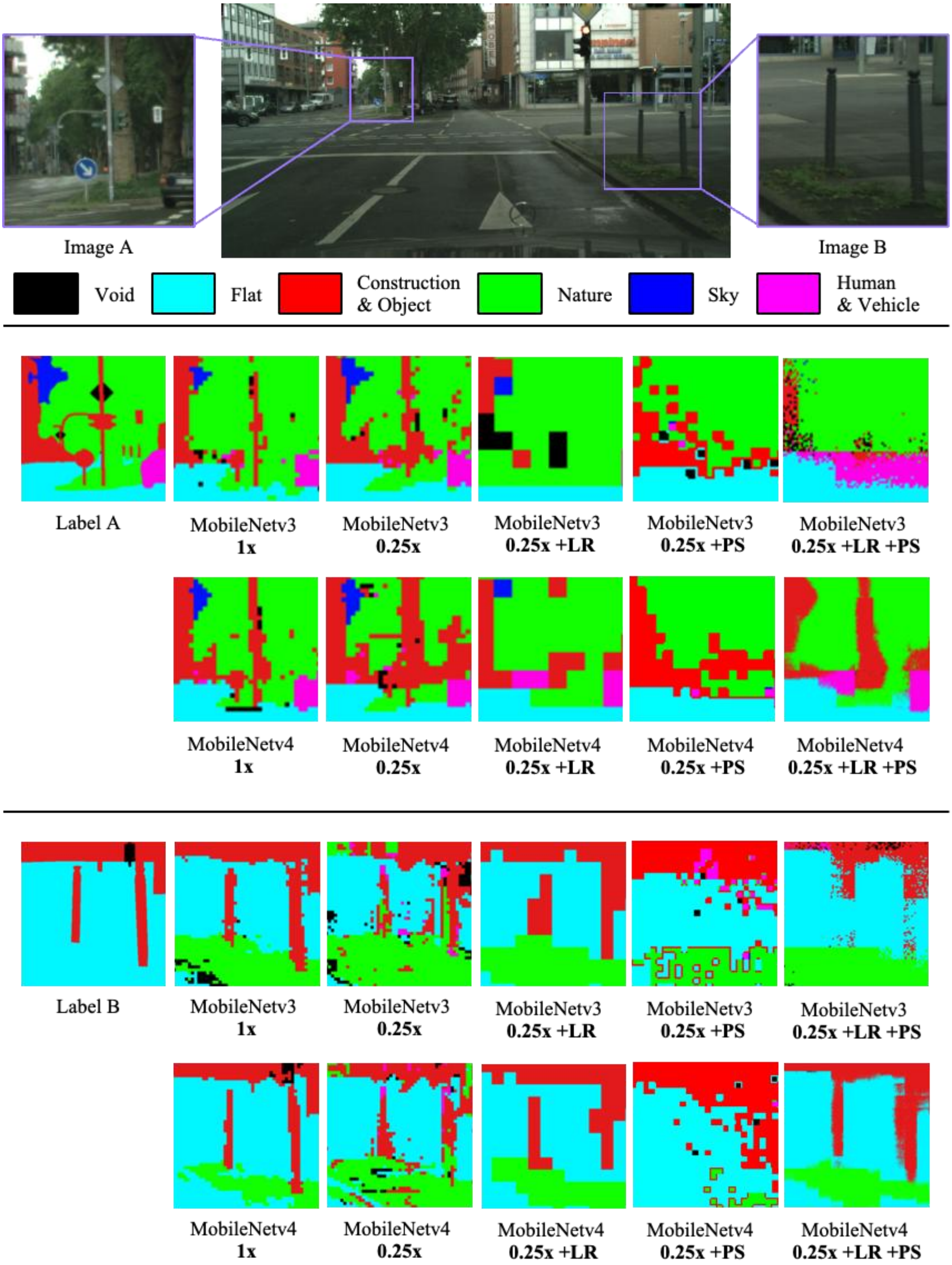

Figure 3. Sample segmentation results highlighting thin objects at varying distances.

A qualitative analysis of sample segmentation results in Figure 2 shows that while the numerical accuracy difference between MobileNetV3 and MobileNetV4 is relatively small, their segmentation outputs revealed key distinctions. MobileNetV4 produced more coherent and better clustered segmentations, effectively reducing large, misclassified regions, such as those belonging to the void class. This suggests that, beyond numerical metrics, MobileNetV4 offers better spatial consistency, making it more effective at distinguishing between differences between regions of complex urban scenes. However, when the images are downsampled, both models experience a notable increase in misclassifications, making their differences less apparent and highlights the challenge of preserving features at lower resolutions.

The impact of the proposed resizing techniques is also evident in their individual qualitative results. Learnable resizing alone tends to produce large, blocky, and well-clustered segmentations, maintaining broad spatial groupings between categories. In comparison, pixel shuffling alone produces more highly irregular boundaries and higher misclassifications within groupings and along the boundaries. When combined, learnable resizing and pixel shuffling complement each other: learnable resizing maintains overall clustering, while pixel shuffling restores spatial details and reduces noise-like misclassifications. This suggests that while learnable resizing alone may suffice for tasks involving general grouping, the combination is preferable for tasks requiring more precise boundary delineation and better structural preservation.

A closer examination of sample segmentation results in Figure 3 further highlights these differences and the complementary roles of these two techniques across both MobileNets. Despite the enhancement provided by learnable resizing and pixel shuffling, both models still struggle to accurately segment thin, pole-like objects at a distance. MobileNetV3 continues to struggle even at shorter distances, whereas MobileNetV4 demonstrates better segmentation in closer-range scenarios. Additionally, while pixel shuffling can restore details in principle, large steps in downsampling reduces its effectiveness for extremely thin or distant objects, sometimes resulting in their complete loss.

Finally, certain misclassifications bring focus to the challenges in dataset labelling. For instance, in Image B a portion of the ground were incorrectly labelled as vegetation, despite being annotated as as flat. Closer inspection revealed the presence of grass in these regions, making the misclassification more understandable. This highlights the importance of carefully inspecting datasets, even when using widely adopted benchmarks like Cityscapes, ambiguities or labeling inconsistencies can influence both model evaluation and performance.

4. Conclusion

This study explored enhancements to real-time urban segmentation by integrating learnable resizing and pixel shuffling into MobileNets. Our findings demonstrate that while our propsal offers improved segmentation accuracy, it initially comes at the cost of slower inference speed.

The learnable resizer proved effective in preserving critical features during downsampling, reducing the accuracy loss typically associated with traditional resizing methods. Meanwhile, pixel shuffling enhanced reconstruction during upsampling, improving segmentation consistency while maintaining a fast inference speed. Together, these techniques helped reduce misclassifications, particularly in complex urban scenes where fine details, such as pole-like structures. With these enhancements, MobileNetV4 outperforms MobileNetV3 in accuracy and greatly reduces the gap between their inference speeds, effectively balancing the trade-off between performance and efficiency. Moreover, precision and recall analyses confirm stable predictive performance, and qualitative results show improved spatial consistency and detail preservation.

Overall, our study highlights the potential of combining lightweight architectures with learnable resizing and upsampling techniques to enhance segmentation performance in resource-constrained environments. By further optimizing MobileNets for real-time urban analysis, we provided a practical solution for applications such as environmental monitoring, autonomous navigation, and smart city planning.

Future work could explore further refinements, including additional attention mechanisms to both the learnable resizing and pixel shuffling, to further enhance segmentation accuracy and prediction speeds. Additional evaluations on additional datasets or real-world camera feeds are also needed to be done to confirm generalizability and assess robustness of the proposal.

References

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016: The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 3213–3223.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019: Searching for MobileNetV3. *Proc. Int. Conf. Comput. Vis. (ICCV), 1314–1323.*

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., Howard, A., 2024: MobileNetV4 – Universal Models for the Mobile Ecosystem. *Proc. Eur. Conf. Comput. Vis. (ECCV), 78-96.*

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1874–1883.

Talebi, H., Milanfar, P., 2021: Learning to Resize Images for Computer Vision Tasks. *Proc. Int. Conf. Comput. Vis. (ICCV), 487–496.*